

Description of GEO supplementary files for “Massively parallel phenotyping of coding variants in cancer with Perturb-seq”

Experimental design

To map variant impact, we profiled A549 cells (ATCC CCL-185) with 64 channels of 10X Chromium Single Cell 3' RNA-seq v2 (10X Genomics #120237), expressing a pool of barcoded variant sequences. The 64 channels were equally split into two experiments: 32 channels for TP53 variants and 32 channels for KRAS variants. We loaded 7,000 cells per channel for each cDNA library to obtain a total of 224,000 cells per experiment (448,000 cells total across the TP53 and KRAS experiments). For each experiment, paired-end libraries were sequenced over 32 lanes on an Illumina HiSeq 2500 per sequencing parameters recommended by 10X Genomics: cell barcode read length 26 bp, index read length 8 bp and transcript read length 98 bp. No dial-out PCR was done, in contrast to typical Perturb-seq (Dixit et al. 2016; Jaitin et al. 2016; Adamson et al. 2016). Variants were assigned to cells based on reads from the RNA-seq data.

File sets

We provide two sets of data analysis files, for each of TP53 and KRAS:

- [rawcounts](#): raw counts from the TP53 and KRAS experiments, after merging cells across the 32 10x channels profiled for each gene.
- [processed](#): processed data from the TP53 and KRAS experiments, as described in the later sections.

Each of TP53 and KRAS experiments also comes with a file that specifies which cell overexpresses which variant, named [A549_<TP53/KRAS>.variants2cell.csv](#).

Raw counts data

Processing steps: To filter out low-quality cells and keep the most informative genes, we removed cells with <200 genes/cell, and then removed genes present in less than 3 of the remaining cells. We then further filtered out cells with fewer than 7,000 UMIs/cell and those with a percent mitochondrial UMIs/cell >20%. We are left with the raw UMI counts for each gene in each cell in each experiment. Specific information about each file is below.

- [A549_<TP53/KRAS>.rawcounts.matrix.mtx](#): the raw counts (third column) for each cell (first column) and each gene (second column), in Market Exchange Format (MEX) format. In each row, the first column represents the index of each cell as listed in [A549_<TP53/KRAS>.rawcounts.cells.csv](#). The second column represents the index of each gene as listed in [A549_<TP53/KRAS>.rawcounts.genes.csv](#). The third column represents the raw counts for the cell and gene combination in the current row. Note that the header of the file includes “%%MatrixMarket matrix coordinate

integer general %", followed by a row specifying the total numbers of cells, genes and counts.

- [A549_<TP53/KRAS>.rawcounts.cells.csv](#): the names of the cells in the experiment. The order of the cells matches with the file [A549_<TP53/KRAS>.rawcounts.matrix.mtx](#).
- [A549_<TP53/KRAS>.rawcounts.genes.csv](#): the order of the genes matches with the file [A549_<TP53/KRAS>.rawcounts.matrix.mtx](#).

Map of which cells overexpress which variant

[A549_<TP53/KRAS>.variants2cell.csv](#): annotation of each cell with the variants it expresses and other metadata. Each row represents one cell, with the cell name present as one of the columns (cell). Each cell is also annotated with its batch (defined as the 10X channel it was characterized in), and total counts in the cell (n_counts). Next, there is a column for each variant, specifying the number of UMIs supporting the presence of each variant in each cell. Note: to obtain the normalized expression level of each variant per cell, divide the number of UMIs per variant in a cell by the total counts in the cell (n_counts). Finally, for each cell we provide the variant(s) assigned to it (columns "variant" and "variant.detailed_multi", with "variant" denoting cells with multiple variants as "multiple" and "variant.detailed_multi" listing the multiple variants present in the cell).

Processed data

Processing steps: As in the Raw Counts dataset, to filter out low-quality cells and keep the most informative genes, we removed cells with <200 genes/cell, and then removed genes present in less than 3 of the remaining cells. We then further filtered out cells with fewer than 7,000 UMIs/cell and those with a percent mitochondrial UMIs/cell >20%.

Additionally, we down-sampled cells with >50,000 UMIs to 50,000 UMIs, to avoid cells with unusually high sequencing depth. After filtering and the above down-sampling, to account for any differences in sequencing depth between the 64 10x channel batches, we further down-sampled the UMIs per cell such that all batches would have a median of 20,000 UMIs/cell. For this, we computed the median number of UMIs/cell in each channel, and then down-sampled the UMIs of each cell by a factor defined as the 20,000 desired UMIs/cell divided by the median number of UMIs/cell in the batch of the cell. Specifically, given a cell and this down-sampling factor, we went through each gene and obtained the adjusted number of UMIs by sampling from a binomial distribution with p =down-sampling factor and N =number of UMIs observed for this gene in the cell. This down-sampling procedure adjusted the distributions of UMIs/cell in each batch to have more similar medians. Batches with a median UMI/cell less than 20,000 did not go through this procedure. Note that in practice, similar results are obtained even without this down-sampling. We normalized the expression UMIs per cell to sum to 10,000 in each cell, and then transformed the normalized values to $\log(\text{normalized expression}+1)$ to obtain a raw

expression matrix. We selected variable genes by identifying the genes for which the variance (scaled to a z-score relative to other genes in similar expression bins) exceeded 0.5. We also filtered the variable genes to have raw expression levels between 0.0125 and 4 (Zheng et al. 2017). We regressed out batch (as a discrete {0,1} variable), the number of UMIs/cell, the percent of mitochondrial reads, and the normalized expression of the variant barcode, and converted the resulting residuals to z-scores for each gene across cells. The analyses in the associated manuscript use these z-scores, unless otherwise noted. We performed PCA for dimensionality reduction, keeping the first 50 principal components. We represented the cells in a low dimension using UMAP, using the default values of 15 nearest neighbors per cell and the default minimum distance between embedded points of 0.5. We used the resulting nearest neighbor graph to cluster cells using Louvain clustering (Blondel et al. 2008; Levine et al. 2015). We then subsampled the cells, to obtain 1000 for each variant. Additional information about each file within the processed data is below.

- [A549_<TP53/KRAS>.processed.matrix.mtx](#): the processed counts (third column, as described above) for each cell (first column) and each gene (second column), in Market Exchange Format (MEX) format. In each row, the first column represents the index of each cell as listed in [A549_<TP53/KRAS>.processed.cells.csv](#). The second column represents the index of each gene as listed in [A549_<TP53/KRAS>.processed.genes.csv](#). The third column represents the processed values (z-scores) for the cell and gene combination in the current row. Note that the header of the file includes “%%MatrixMarket matrix coordinate real general %”, followed by a row specifying the total numbers of cells, genes and sum across values.
- [A549_<TP53/KRAS>.processed.cells.csv](#): the names of the cells in the experiment. The order of the cells matches with the file [A549_<TP53/KRAS>.processed.matrix.mtx](#).
- [A549_<TP53/KRAS>.processed.genes.csv](#): the order of the genes matches with the file [A549_<TP53/KRAS>.processed.matrix.mtx](#).
- [A549_<TP53/KRAS>.processed.cells.metadata.csv](#): metadata for each cell. Each row is one cell. Columns include batch, percent mitochondrial reads (percent_mito), total UMIs/cell after downsampling (n_counts_downsampled), total UMIs/cell before downsampling (n_counts_original), the total UMIs from variant barcodes (vbc.counts), Louvain cluster assignment (louvain), cell cycle scores (G1.S, G2.M, M, M.G1, S), cell cycle phase assignment (phase.multi), gene program scores (P0, P1, ...), PC scores (PC0, PC1, ...) and UMAP embedding coordinates (UMAP1, UMAP2)
- [A549_<TP53/KRAS>.processed.genes.metadata.csv](#): metadata for each gene. Each row is one gene. Columns include the number of cells in which the gene was detected (n_cells), whether a gene is highly variable (highly_variable), mean expr (means), dispersion, normalized dispersion, gene program, contribution of gene to PCs (PC0, PC1, ...)